



Data Munging with SPSS Modeler (3)

데이터마이닝을 위한 데이터 – Customer Signature

이 열은 ID 필드로 모든 행들에서 다른 값을 갖는다.
이것은 데이터 마이닝 목적에서는 무시된다.

이 열은 고객 정보 파일에서 왔다.

이 열은 목표 필드로 예측하고자 하는 필드이다.

2610000101	010377	14		A	19.1		14 Spring ...	TRUE
2610000102	103188	7		A	19.1		NULL	TRUE
2610000105	041598	1		B	21.2		71 W. 19 St.	FALSE
2610000171	040296	1		S	38.3		3562 Oak ...	FALSE
2610000182	051990	22		C	56.1		9672 W. 142	FALSE
2610000183	111192	45		C	56.1		NULL	TRUE
2620000107	080891	6		A	19.1		P.O. Box 11	FALSE
2620000108	120398	3		D	10.0		560 Robeson	TRUE
2620000220	022797	2		S	38.3		222 E. 11th	FALSE
2620000221	021797	3		A	19.1		10122 SW 8	FALSE
2620000230	060899	1		S	38.3		NULL	TRUE
2620000231	062099	10		S	38.3		RR 1729	TRUE
2620000300	032894	7		B	21.2		1920 S. 14th	FALSE

이 행은 유효하지 않는
고객 ID를 가지고 있어서,
분석에서 제외되었다.

이 열은 거래 데이터로부터 요약되었다.

이 열들은 참조 테이블에서 가져왔다.
따라서, 이 값들은 여러 번 반복된다.

이 열은 텍스트 필드로 유일한 값을 가진다.
이것은 다른 유도 변수들을 만들기 위해서
사용될 수 있으나, 분석에서는 무시된다.

- ❖ 모든 데이터가 하나의 테이블에 존재해야 한다.
- ❖ 각 행은 기업과 관련 있는 한 개체 (Ex: 고객)에 대응해야 한다.
- ❖ 하나의 값을 갖는 필드는 무시되어야 한다.
- ❖ 대부분이 한 값을 갖는 필드도 가급적 무시되어야 한다.
- ❖ 각 행마다 다른 값들을 가지는 필드는 무시되어야 한다.
- ❖ 예측 모델링을 위해서 목표 필드와 지나치게 높은 상관관계를 갖는 필드는 제거되어야 한다.



파생변수를 생성하는 일반적인 방법

- 한 값으로부터 특징들을 추출한다.
 - 날짜로부터 요일을 계산
 - 신용카드번호로부터 신용카드 발급자를 추출
- 한 레코드 내의 값들을 결합한다.
 - 멤버십 가입일과 첫 구매일로부터 경과를 계산
- 다른 테이블의 부가적인 정보를 참조한다.
 - 우편번호에 따른 인구와 평균가계수입
 - 상품코드에 대한 계층 구조
- 다수 필드 내에 시간 종속적인 데이터를 pivoting한다.
 - 월마다 한 행씩 저장되는 과금 데이터를 각각의 월에 대응하는 필드로 변환
- 거래 레코드들을 요약한다.
 - 연간 총 구매액
- Customer Signature 필드들을 요약한다.
 - 값의 표준화 및 서열화

고객데이터 - 참조데이터 병합 & 정제 (1/2)

< Customers >

Field	Filter
custid	→
sex	→
birth	→
birth_flg	→
card_flg1	→
mrg_date	→
mrg_flg	→
h_type1	→
h_type2	→
hobby	→
job_type	→
income_flg	✗
car_type	✗
rel_type	✗
dmnot_flg	✗
tmnot_flg	✗
pur_date1	✗
pur_date2	✗
pur_date3	✗
ent_date	→
card_flg2	✗
mail_flg	→
card_str	→
mail_zip1	→
mail_zip2	→
home_zip1	→
home_zip2	→
work_zip1	→
work_zip2	→
cus_stype	→
ncus_stype	✗
m_str1	→
m_time1	→
autopat	→
billnot_flg	✗
fml_cnt	→

< Card_flg >

Field	Filter
card_flg_cd	→
card_sflg_nm	→
card_mflg_cd	✗
card_mflg_nm	✗

< Job_type >

Field	Filter
job_stype_cd	→
job_stype_nm	→
job_gtype_cd	✗
job_gtype_nm	→

< Zipcode DB >

Field	Filter
우편번호	→
일련번호	✗
시도	→
시군구	→
읍면동	→
리	✗
도서	✗
산번지	✗
시작주번지	✗
시작부번지	✗
마침주번지	✗
마침부번지	✗
다량배달처	✗
시작동	✗
마침동	✗
변경일	✗

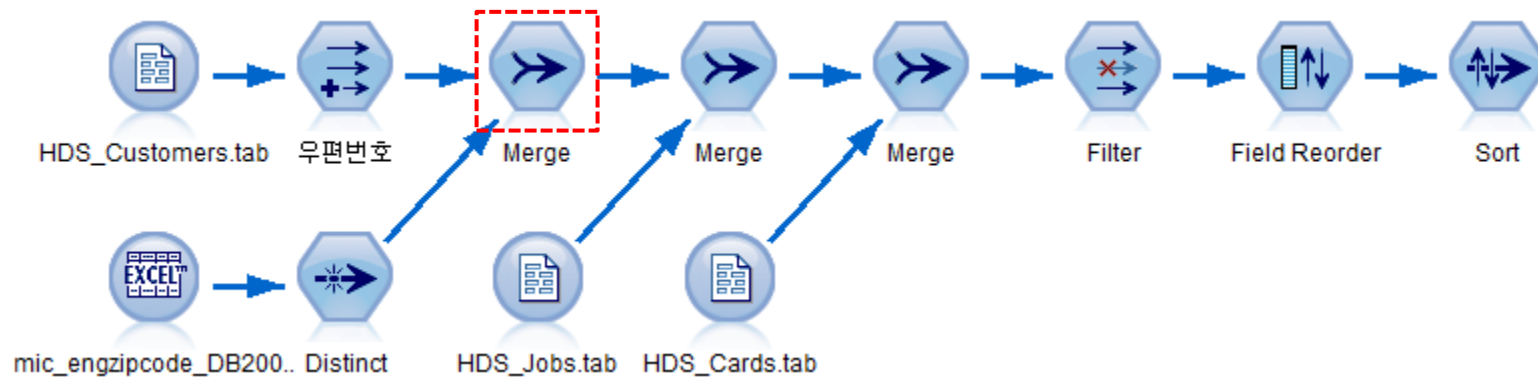
Merge

Merge

Merge

고객데이터 - 참조데이터 병합 & 정제 (2/2)

■ 전체 스트림



■ 유의사항

- ✓ 고객데이터와 우편데이터를 병합할 때 아래의 옵션을 선택해야 함.

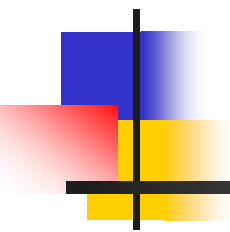
"Include matching and selected non-matching records (partial outer join)"



파생변수 - 선택과 병합

개인과제#1을 통해 수강생이 제안한 파생변수 중 의미 있는 것을 선택하여
고객데이터와 병합하기

- 가격 선호도 변수
- 시즌 선호도 변수
- 상품별 구매 금액/횟수/여부 변수
- 상품별 구매순서 변수
- 주 구매상품 변수
- 휴면/이탈 가망 변수
 - Ex) If 평균구매주기 < 최종구매경과(현재 - 마지막 구매시점) Then “이탈”



ARM (Association Rule Mining, 연관규칙탐사)

연관규칙탐사(ARM)란 ?

- 연관규칙탐사(ARM: Association Rule Mining) : 하나의 거래나 사건에 포함되어 있는 항목들의 경향을 파악해서 상호 연관성을 발견 하는 것
EX) Products in Shopping Cart (One trip, Together)



- 1) 구매자가 제품을 구매할 때 이웃의 영향이 있었는가?
- 2) 오렌지 주스와 청정재 구입시 윈도우 클리너를 같이 구입하는가?
- 3) 우유를 바나나 구입시 함께 구입하는가? 또한 구입 할 때 특정 브랜드를 구입 하는가?
- 4) 청정재를 어느 곳에 위치시켜야지만 판매고를 최대화하는가?



연관규칙(Association Rule) (1/3)

- 어떤 Item 집합의 존재가 다른 Item 집합의 존재를 암시하는 것을 의미하며 다음과 같이 표시한다.

(Item set A) \Rightarrow (Item set B)

(if A then B : 만일 A 가 일어나면 B 가 일어난다.)

- 함께 구매하는 상품의 조합이나 서비스 패턴 발견하는데 이용
- 특정 제품 또는 사건들이 동시에 발생 하는 패턴을 파악하는데 이용
EX) 가정 용품 판매 기간 동안 같이 판매해야 하는 상품의 패턴 발견

연관규칙(Association Rule) (2/3)

Buying Pattern



전항(Antecedent)



야채



생선

후항(Consequent)



포도주? 맥주?

전항(Antecedent)



핸드폰



책 연결기

후항(Consequent)



이어폰? 메모리?

결론 \leq 전제(1) & 전제(2) & ... & 전제(m)
Consequent Antecedents

연관규칙(Association Rule) (3/3)



Pattern Miner System



연관규칙의 평가기준 (1/3)

■ 지지도 (Support)

- 전체 거래 중 항목 X와 항목 Y를 동시에 포함하는 거래가 어느 정도인가 ?

$$S = P(X \cap Y) = \frac{\text{품목X와 품목Y를 포함하는 거래 수}}{\text{전체 거래 수}(N)}$$

- 전체적 구매도에 대한 경향을 파악



연관규칙의 평가기준 (2/3)

■ 신뢰도 (Confidence)

- 항목 X를 포함하는 거래 중에서 항목 Y가 포함될 확률은 어느 정도인가 ?

$$C = P(Y | X) = \frac{P(X \cap Y)}{P(X)}$$

$$= \frac{\text{품목X와 품목Y를 포함하는 거래 수}}{\text{품목X를 포함한 거래 수}}$$

- 조건부확률
- 연관성의 정도
- not symmetric



연관규칙의 평가기준 (3/3)

■ 향상도 (Lift)

- 항목 X를 구매한 경우 그 거래가 항목 Y를 포함하는 경우와 항목 Y가 X와 무관하게 임의로 구매되는 경우의 비율

$$L = \frac{P(Y | X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)}$$

Lift	의 미	예
1	두 품목이 서로 독립적인 관계	과자와 후추
> 1	두 품목이 서로 양의 상관 관계	빵과 버터
< 1	두 품목이 서로 음의 상관 관계	지사제, 변비약

연관규칙탐사 예제

고객의 구매 상품 List

ID	판매 상품
1	소주 , 콜라 , 맥주
2	소주 , 콜라 , 와인
3	소주 , 주스
4	콜라 , 맥주
5	소주 , 콜라 , 맥주 , 와인
6	주스

지지도가 50% 이상인 연관성 규칙

지지도 50% 이상인 규칙	해당 Transaction	신뢰도
소주 => 콜라	1,2,5	75 %
콜라 => 맥주	1,4,5	75 %
맥주 => 콜라	1,4,5	100 %

- Lift = $P(\text{콜라}|\text{맥주}) / P(\text{콜라}) = 1 / (4/6) = 1.5$

*** 연관규칙 : 맥주를 구입한 사람들 모두는(100%) 콜라도 구매한다**

- 지지도: 그리고 이러한 경향을 가지는 사람들은 전체의 절반(50%) 정도이다.
- 리프트: 맥주 구매 시 콜라를 구입하게 될 가능성은 맥주 구매가 전제되지 않았을 경우보다 1.5배나 높아진다.



연관규칙탐사 프로세스

적절한 **Item Set** 결정 및 분석 수준 결정

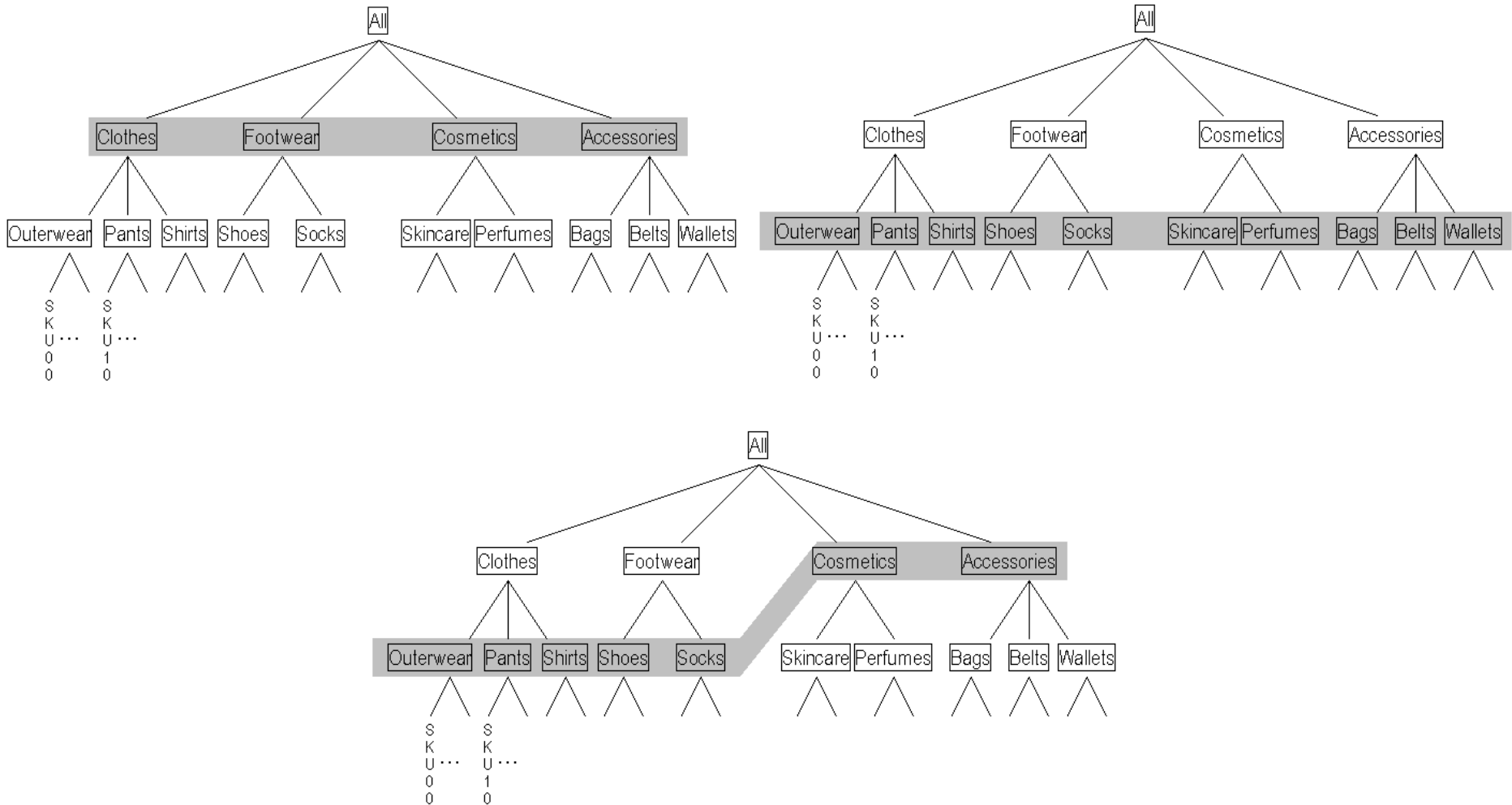
상품간 단순 패턴 발견

- 소주와 콜라, 맥주와 콜라는 타(他)상품의 경우보다 동시구매 횟수가 높다.
- 주스는 맥주, 콜라, 와인과 동시에 구매되지 않는다.

연관규칙 발견

- 지지도, 신뢰도, 리프트 값을 통한 연관규칙의 유용성 분석
- 유용한 연관규칙 결정

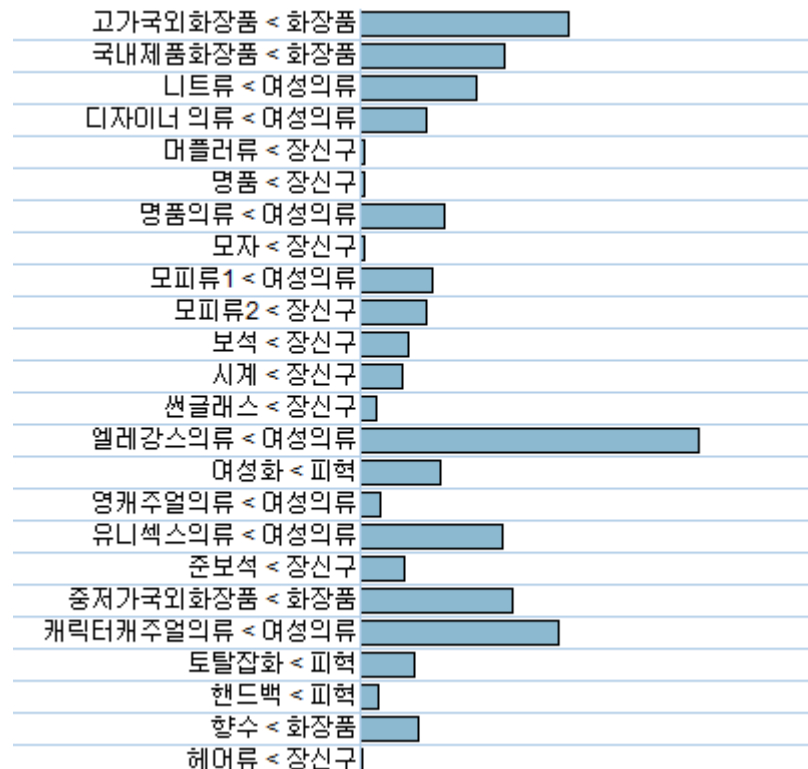
Item 분석수준(Grain) 결정의 예 (1/2)



Item 분석수준(Grain) 결정의 예 (2/2)

H백화점 여성용품 Case

마리끌레르 < 명품의류 < 여성의류
 마리나라날디 < 모피류1 < 여성의류
 마리아주 < 캐릭터캐주얼의류 < 여성의류
 마인 < 유니섹스의류 < 여성의류
 막스마라 < 모피류1 < 여성의류
 막스마라행사 < 모피류1 < 여성의류
 막스앤스펜서 < 명품의류 < 여성의류
 막스앤코 < 모피류1 < 여성의류
 말로 < 모피류1 < 여성의류
 머스트비 < 캐릭터캐주얼의류 < 여성의류
 메세지 < 캐릭터캐주얼의류 < 여성의류
 메세 < 여성화 < 피혁
 메이컬포에버 < 중저가국외화장품 < 화장품
 메키 < 니트류 < 여성의류
 메트로씨티 < 여성화 < 피혁
 모네 < 준보석 < 장신구
 모라도 < 니트류 < 여성의류
 모르간 < 캐릭터캐주얼의류 < 여성의류
 모스키노 < 모피류1 < 여성의류
 미끄마끄 < 명품의류 < 여성의류
 미니멈 < 유니섹스의류 < 여성의류
 미샤 < 유니섹스의류 < 여성의류
 미세스정 < 영캐주얼의류 < 여성의류
 미소니 < 모피류1 < 여성의류
 미소페 < 여성화 < 피혁
 미스로즈 < 니트류 < 여성의류
 미스박 < 디자이너 의류 < 여성의류
 미스식스티 < 엘레강스의류 < 여성의류
 미스지 < 디자이너 의류 < 여성의류
 미쏘니 < 모피류1 < 여성의류
 미쓰제이 < 여성화 < 피혁
 미오르제타 < 모피류2 < 장신구
 미찌코런던 < 여성화 < 피혁
 민수에여성의류 < 유니섹스의류 < 여성의류
 밀라노스토리무역 < 유니섹스의류 < 여성...
 밀로스 < 니트류 < 여성의류
 바닐라 < 캐릭터캐주얼의류 < 여성의류
 바바라 < 니트류 < 여성의류
 바비브라운 < 고가국외화장품 < 화장품
 바이블랙 < 유니섹스의류 < 여성의류
 박동준 < 디자이너 의류 < 여성의류
 박순영니트 < 니트류 < 여성의류
 박윤수 < 디자이너 의류 < 여성의류





연관규칙탐사의 결과유형

■ Useful Result

- 마케팅 전략상 유용한 결과가 나온 경우
- EX) 주말을 위해, 목요일 소매점에 기저귀를 사러 온 아빠들은 맥주도 함께 사 간다. => 주말에 FOOTBALL을 보면서 마심

■ Trivial Result

- 기존의 마케팅 전략에 의해 연관성이 높게 나온 경우
- EX) 정비계약을 맺은 소비자들은 많은 설비를 구매 할 것 같다. => 정비계약은 대개의 경우 따로 맺어지는 것이 아니라, 많은 설비 구입시 함께 제시된다.

■ Inexplicable Result

- 의미를 발견하기 위해 많은 고민이 필요한 경우
- EX) 새로 철물점을 개업하면, 대개 화장실 문고리를 많이 사 간다.



의미 있는 연관규칙의 도출

- 지지도 값의 최소기준치를 미리 설정하여
- 최소기준치 이상의 지지도 값을 갖는 규칙을 생성한다.
- 생성된 규칙 중 높은 신뢰도를 갖는 규칙들을 의미 있는 연관규칙으로 선정한다.
- 자주 구매되는 상품에 대해서 지지도와 신뢰도가 우연히 높게 나올 수 있다.
 - 리프트 (>1)



연관규칙탐사의 장단점

■ 장점

- 명확한 결과 이해
- Undirected Data 분석에 유용
- 다양한 크기의 데이터에 적합
- 신경망이나 유전자 알고리즘에 비해 단순

■ 단점

- 문제의 크기가 커질수록 지수적으로 증가
- 데이터 속성에 대한 제한적 지원
- 항목에 대한 올바른 수 결정의 어려움
- 희박한 항목에 대해서는 문제화



연관규칙탐사 활용 분야

교차판매 (Cross-Selling), 상승판매 (Up-Selling)

- ✓ 스펜서 존슨의 '누가 내 치즈를 옮겼을까?'라는 책을 구매한 고객에게 최인훈의 '상도' 연관 상품을 추천하는 데 활용

부정탐지 (Fraud Detection) : Negative Rule의 활용

- ✓ 신용카드 회사와 같은 금융기관에서는 연관성 규칙을 이용하여 카드 도용과 같은 부정행위를 적발하는 데 활용
 - Negative Rule의 활용
 - Negative Rule은 조건과 결과에 'True' 뿐만 아니라 'False'를 포함한다. 예) $\sim A \Rightarrow B$, $A \Rightarrow \sim B$, $\sim A \Rightarrow \sim B$ 등

매장의 상품진열 (Shelf Planning)

- ✓ 「케이크 ► 와인」이라는 유용한 연관 규칙이 발견 되었다면, 케이크와 와인 상품을 나란히 진열하여 동시 구매를 유도하는 데 활용



연관규칙탐사 알고리즘

Apriori

- ✓ 최소 규칙 지지도(Support), 최대 규칙 신뢰도(Confidence), 최대 전항값 수(Antecedent)로 규칙 생성
- ✓ 품목필드가 이분형(flag) 또는 범주형(set)인 경우에 적용 가능
- ✓ 음의 규칙(Negative rule) 생성 가능

GRI (Generalized Rule Induction)

- ✓ J measure에 의해 규칙의 가치를 평가 : 지지도와 신뢰도를 동시에 고려

$$J = \Pr(X) * [\Pr(Y | X) * \log \left\{ \frac{\Pr(Y | X)}{\Pr(Y)} + (1 - \Pr(Y | X)) \right\} * \log \frac{1 - \Pr(Y | X)}{1 - \Pr(Y)}]$$

- ✓ 연속형 필드를 전항(Antecedent)에 투입하는 것이 가능
- ✓ 단점 : 계산의 복잡성

Sequence (순차규칙)

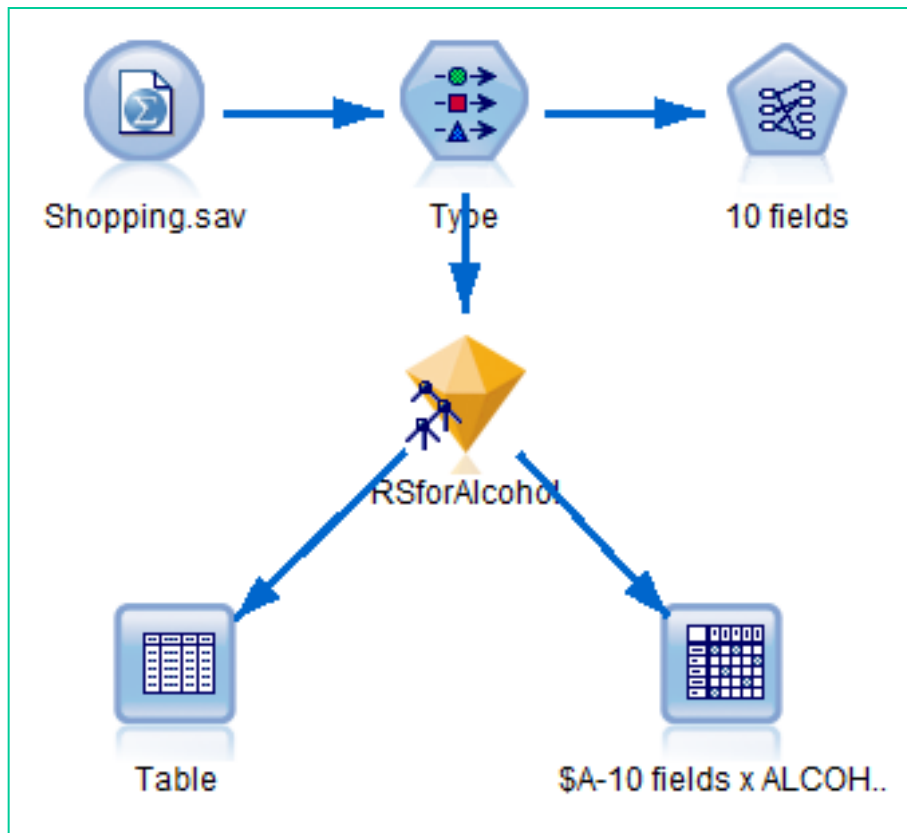
- ✓ 개별 거래 이력에 순차성, 즉 발생시간(time)을 고려
- ✓ 시간필드가 필요
- ✓ 특정한 event가 발생한 이후에 순차규칙이 적용
 - ☞ e.g.) 결혼 후, 유아복 구매하고 교육보험상품을 구매



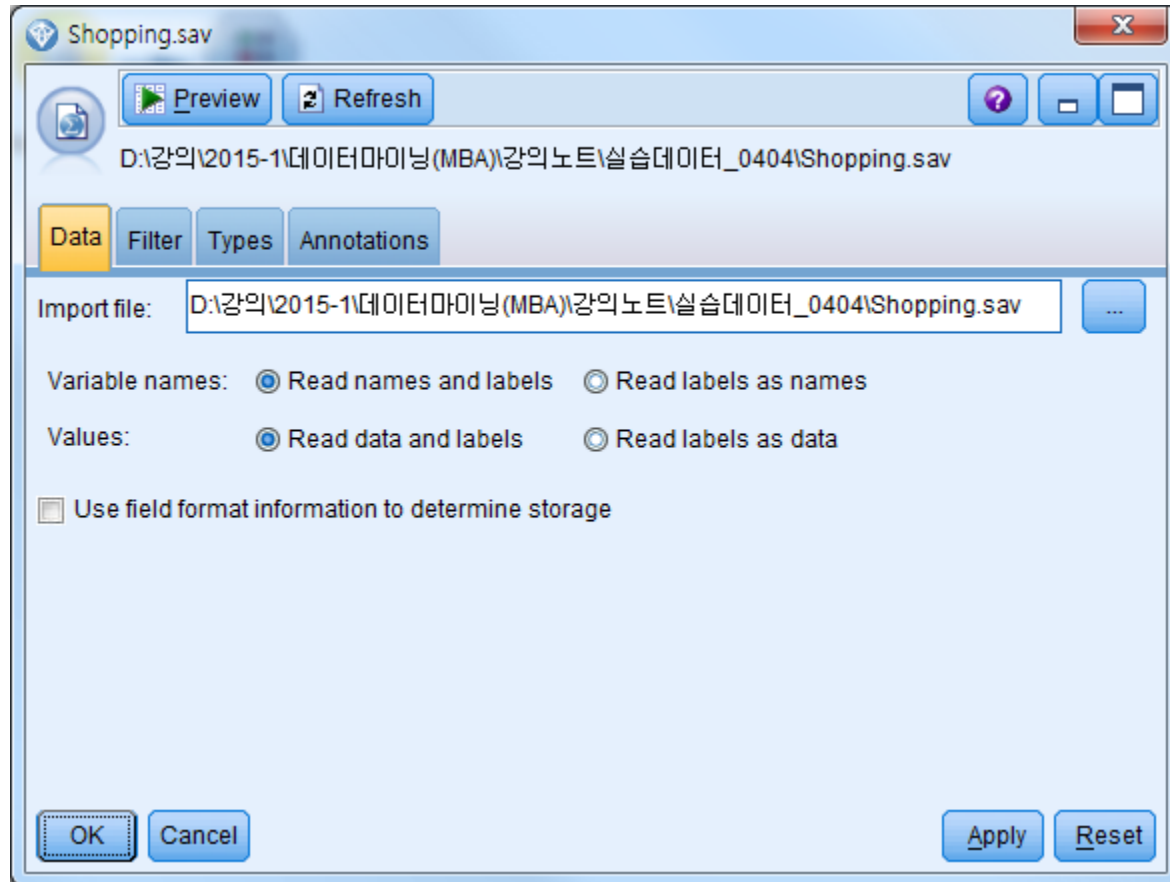
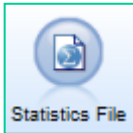
사례 실습: S-Market 장바구니 분석

- Shopping.sav (n = 786 레코드)
- 변수 1-10 (구입품목) 및 분포
 - Ready made (0: 51%, 1: 49%)
 - Alcohol (0: 61%, 1: 39%)
 - Milk (0: 81%, 1: 19%)
 - Fresh meat (0: 97%, 1: 3%)
 - Snacks (0: 52%, 1: 48%)
 - Frozen foods (0: 60%, 1: 40%)
 - Fresh vegetables (0: 92%, 1: 8%)
 - Bakery goods (0: 57%, 1: 43%)
 - Toiletries (0: 90%, 1: 10%)
 - Tinned goods (0: 54%, 1: 46%)
- 변수 11-15 (인구사회적 변인) 및 분포
 - Gender (Female: 54%, Male: 46%)
 - Age (18-30: 30%, 31-40: 25%, 41-50: 17%, 51-60: 16%, 61+: 12%)
 - Marital (Single: 25%, married: 24%, separated: 19%, Widowed: 19%, Divorced 13%)
 - Children (No 65%, Yes 35%)
 - Working (No 17%, Yes 83%)

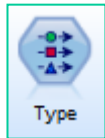
SPSS Modeler의 연관성규칙: 전체 스트림



Statistics File 노트



Type 노드



Type

Preview

Types Format Annotations

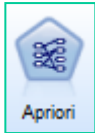
Read Values Clear Values Clear All Values

Field	Measurement	Values	Missing	Check	Role
# READMADE	Flag	1.0/0.0		None	Both
# FROZEN	Flag	1.0/0.0		None	Both
# ALCOHOL	Flag	1.0/0.0		None	Both
# VEG	Flag	1.0/0.0		None	Both
# MILK	Flag	1.0/0.0		None	Both
# BAKERY	Flag	1.0/0.0		None	Both
# MEAT	Flag	1.0/0.0		None	Both
# TOILETRY	Flag	1.0/0.0		None	Both
# SNACKS	Flag	1.0/0.0		None	Both
# TINNED	Flag	1.0/0.0		None	Both
# GENDER	Continuous	[0.0, 1.0]		None	None
# AGEGROUP	Continuous	[1.0, 5.0]		None	None
# MARITAL	Continuous	[1.0, 5.0]		None	None
# CHILDREN	Continuous	[0.0, 1.0]		None	None
# WORKING	Continuous	[0.0, 1.0]		None	None

☒ View current fields ☐ View unused field settings

OK Cancel Apply Reset

Apriori 노드



10 fields

Fields Model Expert Annotations

Model name: ☐ Auto ☐ Custom

☒ Use partitioned data

Minimum antecedent support (%):

Minimum rule confidence (%):

Maximum number of antecedents:

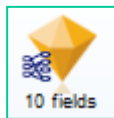
☒ Only true values for flags

Optimize: ☒ Speed ☐ Memory

OK Run Cancel Apply Reset

최소 지지도=0.1
최소 신뢰도=0.75
최대 전항수=5
Positive Rule만 생성

Apriori 노드의 실행결과 (1/3)



- Add To Stream
- Browse** ✓
- Rename and Annotate
- Generate Modeling Node
- Save Model
- Save Model As...
- Store Model...
- Export PMML...
- Add to Project
- ✗ Delete
- Delete

10 fields

File Generate

Model Summary Annotations

Sort by: Confidence % 26 of 26

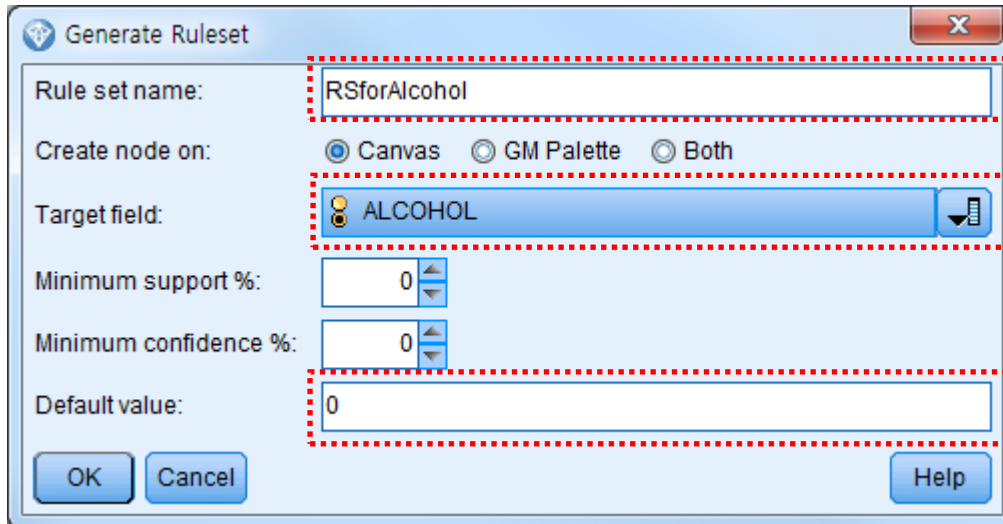
Consequent	Antecedent	Support %	Confidence %
BAKERY	MILK FROZEN	10.814	83.529
BAKERY	ALCOHOL TINNED READMADE	12.087	83.568
BAKERY	FROZEN TINNED SNACKS	11.45	82.222
READMADE	ALCOHOL TINNED BAKERY	12.341	81.443
READMADE	ALCOHOL TINNED SNACKS	11.578	79.121
BAKERY	MILK READMADE	13.359	79.048
BAKERY	MILK TINNED	12.723	79.0
BAKERY	MILK ALCOHOL	11.45	78.889
BAKERY	FROZEN TINNED READMADE	12.595	78.788
READMADE	MILK TINNED	10.051	78.481

OK

- Generate
- Generate Modeling Node
 - Model to Palette
 - Select Node
 - Rule set..** ✓
 - Filtered Model

IF Milk & Frozen
THEN Bakery
(지지도=10.8%
신뢰도=83.5%)

Apriori 노드의 실행결과 (2/3)



Ruleset 생성
(For Alcohol)



Type 노드에
연결

Rule Sets for Alcohol

- (1) Alcohol ≤ Frozen foods & Milk(85: 10.8%, 0.776)
- (2) Alcohol ≤ Ready made & Snacks & Frozen foods(100:12.7%, 0.76)
- (3) Alcohol ≤ Ready made & Bakery goods & Frozen foods(109:13.9%, 0.752)

규칙 1)은 Frozen foods, Milk를
구입한 총 85명의 고객에 적용되
는 것으로 이 규칙의 신뢰도는
77.6%임
-> 전체 고객 중 알코올 구입 비율
이 39%인 것과 비교해 볼 때 상당
히 정확한 규칙이라 할 수 있음

Apriori 노드의 실행결과 (3/3)

- Alcohol에 대한 Rule Set 생성결과: Table 노드 실행

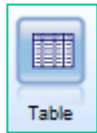


Table (17 fields, 786 records)

File Edit Generate

Table Annotations

	ED	GEND...	AGEG...	MARITAL	CHI...	WORKI...	\$A-10 fields	\$AC-10 fields
1	0	0	1.00	4.00	0	1	0.000	0.500
2	0	0	1.00	3.00	0	1	0.000	0.500
3	0	1	1.00	1.00	0	1	0.000	0.500
4	0	0	1.00	4.00	0	1	0.000	0.500
5	0	0	1.00	3.00	0	1	0.000	0.500
6	1	1	1.00	1.00	0	0	0.000	0.500
7	0	0	1.00	1.00	0	0	0.000	0.500
8	0	0	1.00	4.00	0	0	0.000	0.500
9	1	0	1.00	1.00	0	0	0.000	0.500
10	0	0	1.00	1.00	0	0	0.000	0.500
11	0	0	1.00	1.00	0	0	0.000	0.500
12	0	0	1.00	1.00	0	0	1.000	0.763
13	0	1	1.00	4.00	0	0	0.000	0.500
14	1	0	1.00	1.00	0	0	0.000	0.500
15	0	1	1.00	1.00	0	0	0.000	0.500
16	0	1	1.00	1.00	0	0	0.000	0.500
17	0	1	1.00	1.00	0	0	0.000	0.500
18	0	0	1.00	1.00	0	0	0.000	0.500
19	1	1	1.00	1.00	0	0	0.000	0.500
20	1	0	1.00	1.00	0	0	0.000	0.500

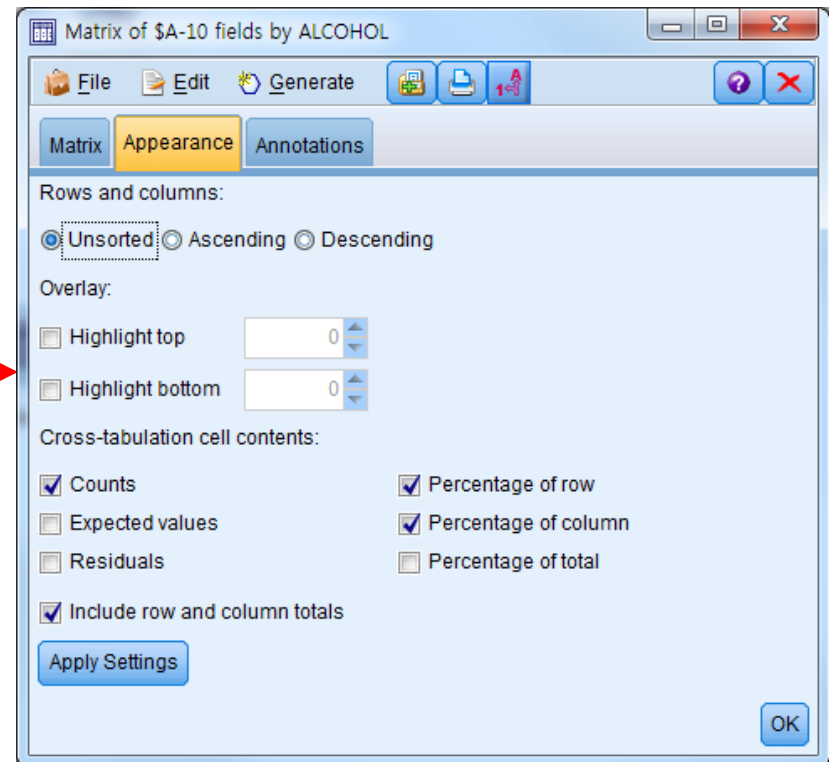
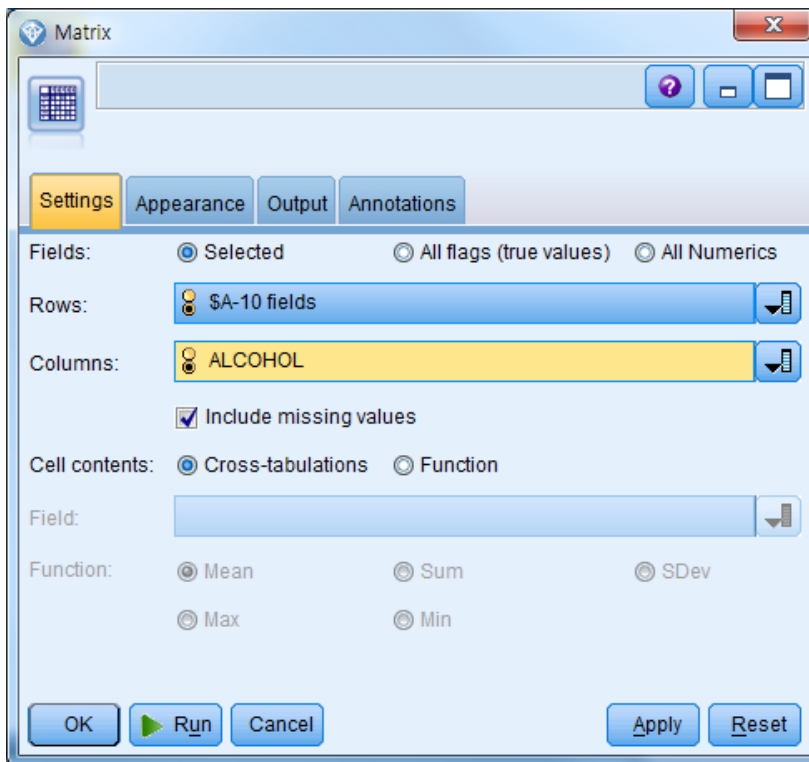
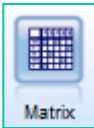
1

OK

12번 고객이
Alcohol을 구매
할 것으로 예측
(확률=76.3%)

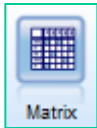
평가 (예측 대 실제) (1/2)

- Alcohol에 대한 Rule Set 평가: Matrix 노드 설정



평가 (예측 대 실제) (2/2)

- Alcohol에 대한 Rule Set 평가: Matrix 노드 실행결과



Matrix of \$A-10 fields by ALCOHOL

File Edit Generate

Matrix Appearance Annotations

ALCOHOL

\$A-10 fields		0.0	1.0	Total
0.0	Count	429	194	623
	Row %	68.860	31.140	100
	Column %	90.126	62.581	79.262
1.0	Count	47	116	163
	Row %	28.834	71.166	100
	Column %	9.874	37.419	20.738
Total	Count	476	310	786
	Row %	60.560	39.440	100
	Column %	100	100	100

Cells contain: cross-tabulation of fields (including missing values)

Chi-square = 86.659, df = 1, probability = 0

OK

ARM Rule Set의 정확도는 69.3%